Biterm Pseudo Document Topic Model for Short Text

Lan Jiang, Hengyang Lu, Ming Xu and Chongjun Wang
National Key Laboratory for Novel Software Technology at Nanjing University
Department of Computer Science and Technology
Nanjing University, Nanjing, China
jianglan1350@gmail.com, hylu@smail.nju.edu.cn, xuming0830@gmail.com, chjwang@nju.edu.cn

***Abstract*—In the past few years, we have witnessed a rapid development of online social media, from which we can access various short texts. Understanding the topic patterns of these short text is significant. Traditional topic models, like LDA, are not suitable when applied to short text topic analysis due to data sparsity. A lot of efforts have been made to solve this problem. However, there is still significant space to improve the effectiveness of these short text specific methods. In this paper, we proposed a novel word co-occurrence network based method, referred to as *biterm pseudo document topic model (BPDTM)*, which extended the previous biterm topic model(BTM) for short text. We utilized the word co-occurrence network to construct biterm pseudo documents. The proposed model is promising since it represents words with their semantic adjacent biterms and is able to model the corpus-level semantic relation between two words. Besides, BPDTM naturally lengthens the documents, which alleviate the influence for performance exerted by data sparsity. Experiments demonstrated that our model outperformed two baselines, i.e. LDA and BTM, which proved its effectiveness on short text topic modeling task.**

I. Introduction

We have witnessed various short text data in the real life, including news titles, facebook posts and website titles. Short text is prevalent in different forms and has promising application prospect, which deserved more academic attention upon them. Semantic topic detection is an interesting issue in short text analysis field. Results of topic detection are basic for many interesting applications, e.g. personalized recommendation and text abstraction. However, extracting topics from short text corpora is a challenging task since there are no enough words in it.

A large number of works have been done for topic detection. PLSA and LDA are two commonly admitted models for discovering hidden topics. In these models, each document is considered as a topic vector generated from a certain multinomial distribution while each topic is represented as a word vector generated from a certain multinomial distribution. The multinomial distributions are estimated using maximum likelihood methods. Since words under a certain topic are susceptible to hold similar semantic meaning, conventional topic models are actually trying to uncover the implicit semantic links between words. These models generate words of a single document under a fixed document-topic distribution. Hence they are quite effective to normal corpora where each document contains a large number of words. However, lacking enough words in short text scenario makes it hard for these models to learn semantic relations between words. Therefore, directly employing them to short text corpora is improper.

In this paper, we propose a novel approach based on word co-occurrence network. By utilizing the co-occurrence network, we manipulate pseudo biterm documents which are used to represent words in the vocabulary of the original corpus. Similar to BTM, we explicitly extract the semantic word relations by means of biterms as well. The main idea behind our work is two-fold. 1) We construct the pseudo documents by using ***triangle relations*** among words. Arbitrarily given three nodes in the network, we denote there is a triangle relation among the three words if every two of the three nodes have an edge with each other. The idea derives from the thought that words belong to a triangle are susceptible to be close to others in the relation. 2) Since the adjacent words list of a word is somewhat semantically related to itself, it is comprehensive to represent the topic distribution of word *n* by using the topic distribution over pseudo biterm document constructed by the adjacent word list. Specifically, we first construct the word co-occurrence network, where each node represents a word in the vocabulary and each edge stands for the co-occurrence relation between its two ends. We simply denote weight of the edges as co-occurrence count. For each word we manipulate a pseudo biterm document according to the triangle relations in its adjacent sub-network. Then topics distribution over pseudo corpus can be obtained by employing standard BTM algorithm on the corpus. This distribution is equivalent to word-topic distribution of the original corpus.

Compared to the traditional topic models, the advantages of our approach lie in two aspects, 1) The pseudo documents derived from triangle relations of the word co-occurrence network can bridge those semantically related words that do not appear in same documents. Fig. 2 shows an example of it. 2) Since a pseudo document is constructed for a word with its adjacent word list, its average length is longer than that in original corpus, which to some extend relieve the sparsity of the corpus. We conducted experiments on two real-world datasets. Experiment results prove that our method is superior to the baselines in terms of topic coherence and several quantitative metrics.

The major contributions of this paper are as follow,

- We combine the word co-occurrence network with biterm topic model and construct more semantic meaningful pseudo documents with the help of *word triangle relations* for topic patterns recognition.
- We denote a word as the biterms extracted from the adjacent words list of a word. By doing this, each word is influenced by multiple semantically close words, which renders it less biased in terms of semantics.
- We conduct several experiments on two real-world short text datasets to determine the effectiveness of our method.

The rest part of the paper is organized as follow: in Section 2, we will briefly review the related work. Our approach is discussed in detail in Section 3 and the method for parameters learning follows in Section 4. Experiment results are displayed and contrasted with baselines in Section 5. Finally we will summarize our work in Section 6.

II. Related Work

In the last few years we have witnessed a large amount of works with regard to topic modeling. In this section we briefly introduce typical topic models for normal text, and then summarize various works specific to short text.

Text corpus can be represented as a document-word matrix with the bag-of-word assumption[1]. The problem of topics extraction stemmed from latent semantic indexing (LSI)[2], which decomposes the document-word matrix into smaller matrices by means of singular value decomposition. Despite that LSI succeeded in uncovering topic patterns to some extent, the time-consuming model fails to make an reasonable explanation to negative values in resulting matrices. [3] proposes probabilistic latent semantic indexing (pLSI) which considers a document as a mixture of topics while a topic as a mixture of words. Latent dirichlet allocation (LDA)[4] improves pLSI by imposing Dirichlet priors α and β, which renders the method totally probabilistic. LDA has been widely employed and succeeds in many normal text topic modeling tasks.

A huge amount of extended LDA or pLSI methods have been proposed to solve specific problems, such as social network based topic model[5], author-topic model[6], labeled topic model[7], hierarchy topic model[8]. All these approaches perform well on normal text corpus while none of them has considered the applicability on short text.

The major problem of topic modeling on short text lies in that the document-word matrix is very sparse, which leads the semantic relations between words to be vague. Numerous efforts have been done in solving the problem in the past few years. For instances, [9] aggregated tweets sharing same key words into several lengthy documents and models topics on them. [10] utilized auxiliary long text corpus to learn the topics on short text. [11] extended the traditional non-negative matrix factorization method, which in advance extracted the semantic relations between word pairs. [12] proposed a novel term weighting scheme for NMF, derived from the Normalized Cut problem on the term affinity graph. Some imposed various assumptions to the models. For example, [13] assumed that each document only belonged to a single topic. [14] assumed that each document contained limited topics while each topic could be represented by limited words. [15] proposed a generative biterm topic model which explicitly uncovered the semantic relations between words. [16] extended the biterm topic model, considering it necessary to filter out unimportant biterms that are too general. Among all the methods mentioned above, biterm topic model is closest to our approach. However, none of these approaches directly probe the latent topic distribution via biterm adjacent list extracted from word co-occurrence network.

III. Our Method

Theoretically, LDA-like topic models follow the assumption that words within a single document are generated from a certain multinomial distribution, which derived from the prior Dirichlet distribution. Therefore, results of these models intrinsically hold the word-occurrence patterns in the corpus. However, effectiveness of LDA is seriously influenced when the documents are short, which brings strong sparsity to the training process. On the other hand, words having no explicit co-occurrence can probably have semantic relations as well. To solve both problems mentioned above, we propose a novel topic learning method, in which we construct a pseudo biterm corpus based on the word co-occurrence network. After that, we employ standard BTM to generate the manipulated corpus and infer the topics for documents by using the parameters learned in generative process. Since our model is an extension of BTM, we would like to sketch the processing of BTM in advance first.

A. Biterm Topic Model

Biterm Topic Model (BTM) models the topic patterns of a document corpus based on the biterms which are generated from the corpus. BTM transfers the co-occurrence relation of two words in a document into a biterm first, and construct a corresponding document based on the generated biterms. Finally, BTM models the topics upon the biterm corpus and assumes that both words in a biterm are generated from the same topic-word distribution. Fig. 1 shows the general process of BTM.

B. Word Co-occurrence Network

We construct the word co-occurrence network based on the original corpus. The network is weighted and undirected, of which each node represents a unique word in the vocabulary of original corpus and each edge represents the co-occurrence frequency between its two ends. Obviously, the sub-network concerning every single document in original corpus is a complete graph, which makes the network relatively tight. Words that appear rarely in the original corpus are able to hold more connections with other words.

C. Biterm Pseudo Corpus Generation

Since we are going to employ BTM on the pseudo document corpus, the following work is to extract biterms for the model.

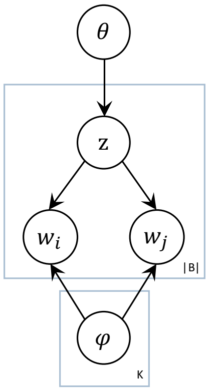


Fig. 1. Graphical representation of BTM. $|B|$ denotes the biterm vocabulary while θ and ϕ represent the document-topic distribution and topic-word distribution respectively

For each node n_0 in the network, we pick out its adjacent sub-network, namely A_0. For each two nodes n_i and n_j in A_0, if there is an edge between them in the original corpus, then we put the biterm $<n_i,n_j>$ inside the pseudo document of n_0 . The commonly-used theory behind our choice is quite intuitive, i.e. 'two friends of mine are probably friends of each other'. Each biterm is possible to appear several times in the pseudo document. We are using the minimum weight of the three edges here since it reduce the overall number of biterms in the pseudo document set without losing the relative weight patterns. Fig. 2 illustrates the procedures of constructing word co-occurrence network and biterm pseudo corpus.

D. Biterm Pseudo Document Topic Model (BPDTM)

With the biterm pseudo corpus in hand, we can now employ the standard biterm topic model to generate the corpus.

Although we use the same generative process to BTM in our method, the intrinsic target and meaning is totally different. BTM learns the parameters which stand for topic-word distribution. Thereby, it directly generates the biterms of the original corpus. In BPDTM, however, the same parameter learning method is used to generate the pseudo document-topic distribution and topic-word distribution. The generative process of BPDTM is depicted in Algorithm 1.

Algorithm 1 Generative process for BPDTM

```
Input: Topics number $K$, Pseudo corpus $\Omega$
Output: $\Phi_z$ ,$\Theta_i$
for each topic z do
    draw a topic-pseudoword distribution $\Phi_z$~Dir($\beta$)
end for
for each pseudo document $d_i$ in the pseudo corpus do
    draw a pseudodocument-topic distribution $\Theta_i$~Dir($\alpha$)
end for
for each biterm $b$ in pseudo document $d_i$ do
    (a) draw a biterm-topic distribution $z_b$~Multi($\Theta_i$)
    (b) for $b$ draw two words $w_i$,$w_j$~Multi($\Phi_z$)
end for
```

E. Document Topics Inference

We can easily infer the topic distribution over the documents in original corpus with the topic distribution over pseudo documents. According to the previous discussion, words distribution over topics is identical to topic distribution over pseudo documents. Therefore, this topic-word distribution can be exploited to derive topic distribution of documents in original corpus. Specifically, we propose that topics of a document can be represented by the expectation of the topics over words within the document, which can be calculated as follow:

$$\begin{aligned} P(z|d) &= \sum_{w_i} P(z|w_i)P(w_i|d) \\ &= \sum_{w_i} P(z|pd_i)P(w_i|d) \end{aligned} \tag{1}$$

where $P(z|pd_i)$ is the topic distribution over pseudo document pd_i that equals topic-word distribution of word w_i we have obtained above and $P(w_i|d)$ is simply the normalized word count in the document, i.e.,

$$P(w_i|d) = \frac{|w_{i,d}|}{|d|} \tag{2}$$

where $w_{i,d}$ stands for the frequency of word i in document d. According to our experiment result, we find it really effective to estimate $P(w_i|d)$ with the normalized frequency, in spite of that it is simple in the form.

Our method is easy to understand and implement. BPDTM explicitly models the semantic relationship between every two words, whether or not they appear in the same documents, which strengthen its ability to uncover the real semantic patterns behind the corpus. Meanwhile, using biterms instead of single word as the unit of document lengthens the documents, which in turn decreases the influences of sparsity.

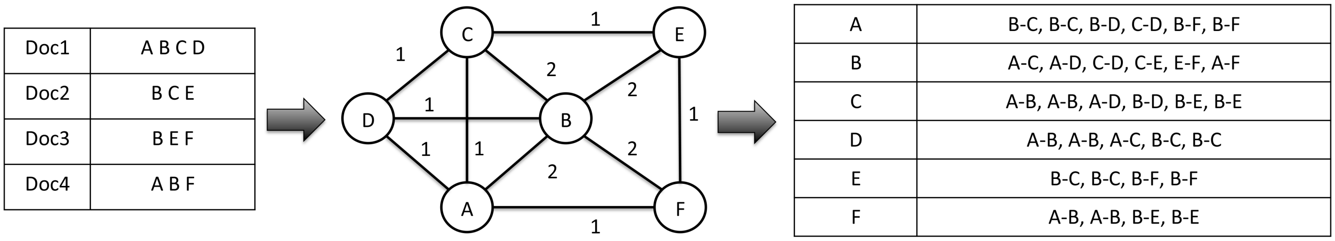
F. Parameter Learning

In BPDTM, we remain to learn the document-topic distribution separately for every document. The feasibility for us to do this comes from that each pseudo document constructed from the adjacent biterms list of the word is long enough. Suppose that there are n documents and the vocabulary size is m. The average documents length is denoted as l. Then for each document, there will be $\binom{l-1}{2}$ biterms, and for the whole corpus $n * l * \binom{l-1}{2}$ biterms. Since the pseudo corpus contains m pseudo documents. The expectation length of each pseudo document $E(l_{pd})$ can be calculated as follow,

$$E(l_{pd}) = \frac{n * l * \binom{l-1}{2}}{m} \tag{3}$$

where l_{pd} denotes the length of a single pseudo document. Notice that equation (3) calculates only the intra-document co-occurrence, the real value will be larger if we take corpus-level co-occurrence into consideration. For BTM, the expected length of each document is $\binom{l}{2}$. Therefore, average pseudo document length is r times larger than average original document length, where,

$$r = \frac{n * l * \binom{l-1}{2}}{m} / \binom{l}{2} = \frac{n}{m}(l-2), \tag{4}$$



Doc1	A B C D
Doc2	B C E
Doc3	B E F
Doc4	A B F

A	B-C, B-C, B-D, C-D, B-F, B-F
B	A-C, A-D, C-D, C-E, E-F, A-F
C	A-B, A-B, A-D, B-D, B-E, B-E
D	A-B, A-B, A-C, B-C, B-C
E	B-C, B-C, B-F, B-F
F	A-B, A-B, B-E, B-E

Fig. 2. Procedures of constructing word co-occurrence network and biterm pseudo corpus. Word *A* and word *E* do not co-occur in original corpus. While after constructing pseudo corpus, they co-occur in the same pseudo documents of *B*, *C* and *F*

in our cases, n is much larger than m, leading that r=22.6 and 21.3 respectively for news title and question title data. With the lengthy pseudo documents, we are able to individually model θ for each pseudo document. The benefit of doing so is clear, that it can avert the strong assumption that all the pseudo documents follow a unique corpus-topic distribution.

We employ Gibbs sampling to learn parameters in our model. The learning procedures is almost the same to that in BTM, except that we learn individual document-topic distribution for each pseudo document.

IV. Time Complexity Analyses and Pseudo Corpus Scale-down

A. Time Complexity Analyses

All the three candidate methods utilize Gibbs sampling to learn parameters. Gibbs sampling considers each unit, which is word in LDA while biterm in BTM and BPDTM, as a single training data and successively upgrades their values in each iteration. Therefore, the time complexity of LDA in one iteration is O($N_d K_z L_d$), where N_d denotes the number of documents and L_d is the average length of documents in the corpus. For all the methods, K_z represents the number of topics. For BTM, the time complexity is O($N_d K_z L_b$), where L_b denotes the average biterm documents length. In BTM, a sliding window is used to limit the scale of biterms. Suppose that the size of sliding window is s, then for each word, $\binom{s}{2}$ biterms can be generated. Normally in short text corpus, s equals the length of document. Therefore, the time complexity of BTM is O($N_d K_z L_d^2$). For BPDTM, time complexity for a single iteration is O($N_p K_z L_p$), where N_p is the number of pseudo documents, i.e. size of the vocabulary and L_p is the average pseudo documents length. From (3) we can see L_p equals $N_d L_d(L_d-1)(L_d-2)/N_p$. Thus, time complexity of BPDTM is O($N_d K_z L_d^3$).

Time complexity of BPDTM is $O(L_d^2)$ times larger than LDA and $O(L_d)$ times larger than BTM, which renders it time-consuming. Therefore, we designed a method to decrease its time consumption.

B. Pseudo Corpus Scale-down

When constructing the pseudo corpus, we took only the co-occurrence of every two words into consideration. However, a number of words in the vocabulary are poor at discriminating topics, which is ignored in the previous work. [16] demonstrated in their work that filtering out weak topic discriminative words could improve the performance of topic clustering. In our work, document frequency is employed to group words into **topical words**(T) and **general words**(G). Document frequency for each word is calculated as follow,

$$f_w = \frac{n_w}{n} \tag{5}$$

where n_w denotes the number of documents that contain word w. Since topical words should appear in a small bunch of documents, we considered words with lower document frequency as topical words. To separate words into topical words and general words, we set a threshold γ and grouped those with document frequency score lower than the threshold into topical words.

After distributing each word into certain group, three biterms type were obtained, which are T-T,T-G and G-G. We then conducted the learning process after removing all the G-G biterms due to our assumption that they are helpless in topic clustering.

V. Experiment

In this section, we will display the experiment results, which demonstrates that our method outperforms another two baselines. The baselines in our experiment are traditional LDA and BTM.

A. Settings

1) Datasets: To verify the effectiveness of our method, we conducted the experiment on two real-world datasets. 1) A Chinese labeled news title dataset published by SogouLab[1] containing part of news title from several Chinese media websites. 2) A question title dataset from a famous Chinese Q&A on-line community[2]. We implemented a specific spider to

[1] http://www.sogou.com/labs/dl/tce.html

[2] http://www.zhihu.com

crawl question titles from the site and label every question title with the hashtags pinpointed by the questioners. Considering that each question contains at least one hashtag, while we actually need only one label for each title, we labeled the title with its most important hashtag, i.e. the first label in its label set.

We employed Hanlp[3], which is an open source Chinese language processing tool, to segment each document into bag of words. After that, we removed stop-words as well as words that appeared less than 6 times and documents that contained no more than 4 words. Additionally, we calculated label entropy for all the remaining words and retained those specific words whose entropy are less than 4. Table I shows the basic information of the two datasets after preprocessing.

TABLE I
BASIC INFORMATION OF THE TWO DATASETS

Datasets	News	Questions
#documents	11536	6022
#words	2946	1460
ave doc length	5.77	5.16
#classes	13	22

2) Environment and Models: All of our experiments were conducted on a Linux server with Intel Core i5 2.9 GHz CPU and 8G Memory. We compared our method with LDA and BTM. We implemented our BPDTM via JAVA. We simply applied the open-source package JGibbLDA[4] for LDA and implementation of BTM from the authors[5]. We empirically set the Dirichlet prior distribution parameters for all the three methods, i.e., $\alpha = 50/K$ and $\beta = 0.1$. The influence of the initial settings is proved later by the result to be negligible. We halted Gibbs sampling after iterating 1000 times and reported the average result of 10 times in order to get robust results.

B. Evaluating quality of topics

Perplexity is a typical measurement for traditional topic models. It measures the results on held-out test subsets. However, it is not suitable for us since both BPDTM and BTM generate corpus indirectly. Instead, we employed topic coherence to evaluate the quality of the resulting topics.

Topic coherence is proposed by Mimno et al. as a way to judge the performance of topic clustering[17]. The main idea lies in that good topic clustering result in cohesive semantic similarity. Therefore, topic coherence calculates the overall semantic similarities of the most representative words in a topic. It is calculated as follow,

$$C(z;V^{(z)}) = \sum_{t=2}^{T}\sum_{l=1}^{t}\log\frac{D(v_t^{(z)},v_l^{(z)})+\varepsilon}{D(v_l^{(z)})} \tag{6}$$

where $V^{(z)} = [v_1^{(z)}, v_2^{(z)}, ..., v_T^{(z)}]$ denotes the words vector that is most representative about the topic z. $D(v_l)$ denotes the word frequency of v_l in the corpus while $D(v_m, v_l)$ is the co-occurrence count of v_m and v_l. The perfect topic partition leads to the topic coherence score approaching to 0. The idea behind topic coherence is that the more a word pair co-occurs, the more possible they belong to the same topic. The empirical practice indicates that results of topic coherence is in accordance with that by intuitive human judgement, which makes it an ideal measurement. ε in Eq.(6) is a tiny positive number used for avoiding zero in the fraction. The reported topic coherence score comes from the average of scores under all topics. We conducted topic clustering task upon the two datasets with all the three algorithms and respectively calculated their topic coherence based on the topic-word distributions. According to the inherent class number, we set K=10 to news titles corpus and K=20 to question titles corpus. The performance is listed in Table II, where the number of most representative words T ranges from 5 to 20. It proves our approach significantly performs better than LDA and BTM (P-value < 0.01 by t-test).

TABLE II
AVERAGE TOPIC COHERENCE SCORE

(a) Average topic coherence for news title

T	5	10	20
LDA	-31.3±0.8	-156.1±2.4	-697.5±3.2
BTM	-28.4±1.6	-147.5±3.1	-658.0±5.7
BPDTM	**-26.9±1.1**	**-130.8±1.5**	**-514.9±4.1**

(b) Average topic coherence for question title

T	5	10	20
LDA	-31.7±0.7	-156.3±1.4	-679.2±3.0
BTM	-26.0±0.3	-131.7±1.1	-599.6±3.2
BPDTM	**-22.4±0.4**	**-107.5±0.9**	**-469.2±3.1**

It is reasonable that BPDTM outperforms LDA on such two datasets, for LDA lacks enough words information under the circumstance of short text and fails to uncover word-word semantic relations at the word-level. Modeling the corpus from word-level relations can improve the topic quality, which is proved by the fact that BTM performs better than LDA. Beside explicitly modeling the relations between words, BPDTM also discovers semantic relations between non-co-occurrence word pairs from corpus-level.

In order to display the most representative words of each topic clearly, we picked out the top 20 words of three randomly chosen topics while K was set as 20. The results were showed in Table III. We displayed the results of the three candidate models respectively. The results of LDA show almost no topic information since we are unable to judge the topic from these words, while BTM and BPDTM both show strong topic patterns. From the topic words we can judge that the three topics are about education, sport and technology. However, BPDTM models the results better, since its most representative words are more topical, which means it contains fewer semantic general words like 'run' in the education topic

[3] http://hanlp.linrunsoft.com/

[4] http://jgibblda.sourceforge.net/

[5] http://code.google.com/p/btm/

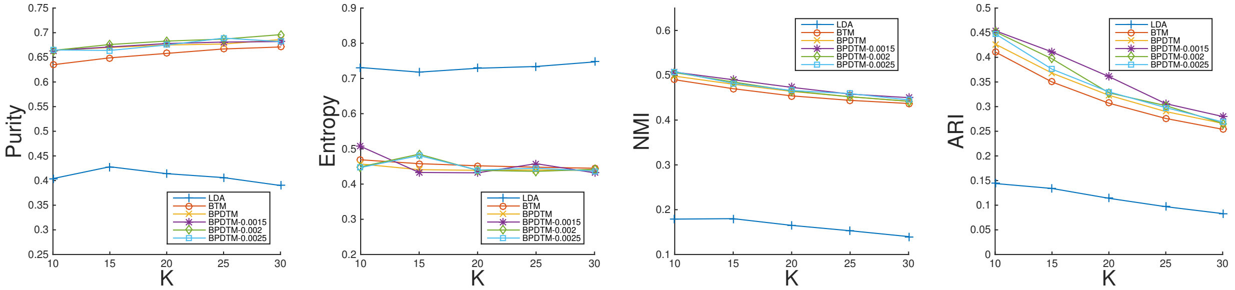


Fig. 3. Comparison of four document clustering metrics on news title corpus

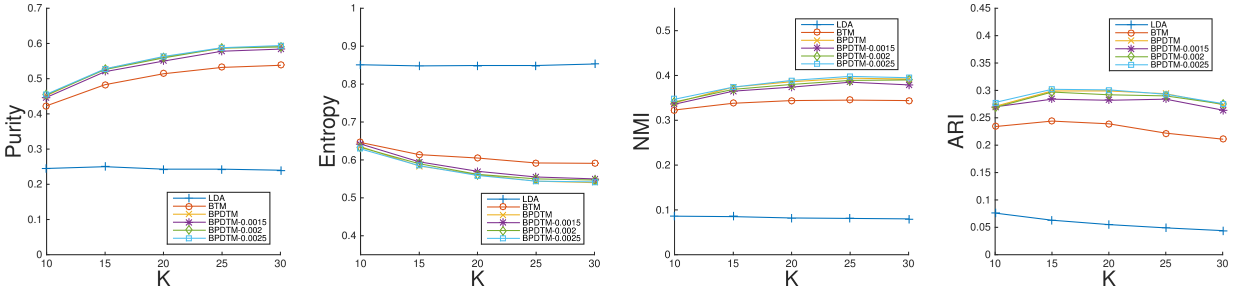


Fig. 4. Comparison of four document clustering metrics on questions title corpus

TABLE III
MOST SEMANTIC REPRESENTATIVE WORDS IN THE TOPICS OF NEWS TITLE CORPUS

LDA	BTM	BPDTM
enroll score beijing stimulate strong HSBC nationalteam decrease controll graduate alibaba area corn house criticism disabled payment degree sell storm	exam unity collegeentranceexam enroll national-wide college English problem university maths paper thesis prediction review answer run beijing politics university biology	collegeentranceexam exam teacher English enroll Chinese maths specialty university score prediction admission biology analysis answer college skill understanding geograpyh IELTS
business chinatelecom purchase charge China basketballteam important traditional chinamobile key rebuild education experience opposite wonderful fee death tell panel score	NBA MilwaukeeBucks yijianlian Lakers finals Kobe player yaoming celts training match rockets prepare basketball season spurs international team CBA netease	Germany fans finals NBA volleyball player champion EuropeanChampionship rockets EuropeanCup Kobe season coach Netherlands match Italy Lakers team training MilwaukeeBucks
google gym navigation arsenal college distribution aid revolution media star individual thinking effectiveness progress anxiety student edge carnival holyflame Lotte	yahoo Microsoft purchase rebuild telecoms unicom business company stock ceo google corporation netcom influence scheme tvb price plan Gates search	Microsoft purchase yahoo rebuild telecoms business department market mobilephone company Chinamobile protocol communication ceo google netcom Gates network Nokia service

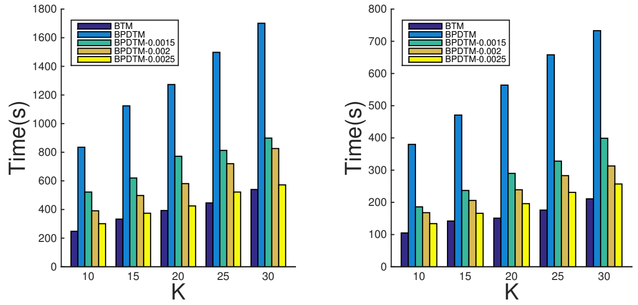
and 'prepare' in the sport topic in the BTM results.

C. Document Clustering

Document clustering is a popular research issue in text analysis. It can partition unlabeled documents into several clusters while documents within a cluster are close to each other in terms of semantics. Document clustering is applicable in various scenarios, e.g. personalized information recommendation and human-computer interaction. Since every document in this task only has one label, we simply label a document with the cluster of highest probability in $P(z|d)$.

Given a labeled dataset $\langle D, C, M\rangle$, in which D is the corpus, $C = [c_1, c_2, ..., c_p]$ is the labels of D as ground truth, and $M = [m_1, m_2, ..., m_k]$ is the clustering result predicted by models. Therefore, the dataset contains p different classes while topic number is set to be k. To verify the effectiveness of our method, we conducted experiments on four different metrics: purity, information entropy[18], normalized mutual information (NMI)[19] and adjusted rand index (ARI)[20].

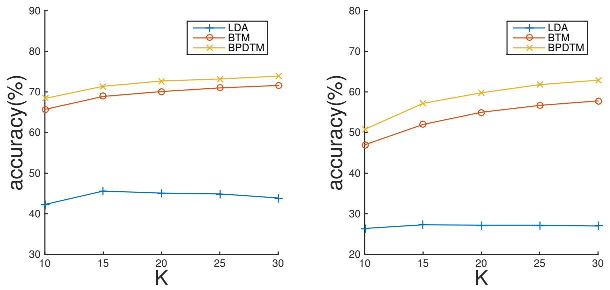
Fig. 3 and Fig. 4 show the resulting scores from our experiments on both the two corpora. We set the cluster number ranging from 10 to 30 with the step size of 5. Results demonstrate that our method outperforms both the baseline methods while BTM performs better than LDA as well. In particular, LDA performs badly compared to both other two methods in these two scenarios, indicating that it is not suitable for short text topic clustering. BPDTM is superior to BTM in all the four metrics, indicating that modeling words via their close biterms lists is useful for topic structure extracting.



(a) News title dataset

(b) Questions title dataset

Fig. 5. Time consumption of 1000 iterations on the datasets.



(a) News title dataset

(b) Questions title dataset

Fig. 6. Classification accuracy on the datasets.

The time consumption of 1000 iterations for BTM and BPDTM are displayed in Fig. 5. For BPDTM, general words threshold γ is utilized to remove general biterms and scales down the corpus to different extents. Empirically, we set γ to be 0.0015,0.002 and 0.0025 in both document clustering and time consumption experiments, which is depicted as $'$BPDTM-0.00x$'$ in the legends. Results show that time consumption of BPDTM drops along with the decrease of γ. On the other hand, clustering performance almost never changes under different threshold, indicating that our pseudo corpus scale-down method is promising to accelerate the learning process.

D. Document Classification

To verify the ability of BPDTM in semantically representing documents, we conduct document classification experiments on both the short datasets.

All of the three approaches in this paper generate a topic distribution for each document d, which can be denoted as follow,

$$d_i = \theta(d) = [p(z_1|d_i), ...p(z_k|d_i)] \tag{7}$$

Therefore, we can represent a document with the topic distribution vector, in which every dimension means its topic proportion of that topic. By doing this, we normalize all the documents with the same dimension size. After representing documents with equation (7), we can consider each document as a vector, of which each dimension is a feature. Consequently, each document maintains a fixed number of features that equals the number of topics. We took advantage of that to do classification. We randomly separated the dataset into training set and test set with the ratio of 9:1 and performed cross validation on them with an open-source classification method LibLINEAR[6]. The detail experiment results are shown in Fig. 6.

It can be observed from the results that both BTM and BPDTM perform greatly better than LDA on the two candidate datasets. Since BTM and BPDTM essentially model the semantic relations between words, we are certainly able to speak that modeling the word-word relations is an effective way to solve the sparse problem in short text topic modeling. Beside that, our BPDTM is better than BTM as well. Two possible reasons for that are, 1) BTM fails to model the semantic relations between word pairs, which do not co-occur while BPDTM does. 2) Unlike BPDTM, BTM does not represent words with their adjacent biterm lists. The advantage of applying the adjacent word-pair list is obvious. It can synthetically take the semantic explaining ability of each adjacent biterm into consideration. On the other hand, since each biterm in the pseudo document is extracted from a triangle relation in the word co-occurrence network, which is indeed a tight relation in terms of semantics, it is reasonable that these biterms are more useful in representing words.

VI. Conclusion

In this paper, we proposed a novel approach BPDTM to obtain precise topics from short text datasets. We employed triangle relations in the word co-occurrence network to construct the pseudo document sets, which were further used to train our algorithm. The idea behind our approach lies in that three words are semantically close if every two of them co-occur in the corpus. We conducted experiments on two real-world datasets, i.e. news title and question title corpuses. Experiment results demonstrated that our approach outperformed the baselines. Additionally, our approach can disclose the word-word relations at the corpus-level, which means it is able to model the semantic relations between words that do not appear together in the corpus. In our future work, we will try to remove unimportant edges in the word co-occurrence network, and subsequently investigate its influence towards topic modeling. Besides, we would like to uncover whether BPDTM can also work well on normal text datasets.

Acknowledgment

This paper is supported by the National Key Research and Development Program of China (Grant No. 2016YFB1001102), the National Natural Science Foundation of China (Grant No. 61375069, 61403156, 61502227) and the Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing University.

References

[1] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2-3, pp. 146–162, 1954.

[6] http://www.csie.ntu.edu.tw/ cjlin/liblinear/

[2] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, p. 391, 1990.

[3] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1999, pp. 50–57.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[5] Y. Cha and J. Cho, "Social-network analysis using topic models," in *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2012, pp. 565–574.

[6] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in *Proceedings of the 20th conference on Uncertainty in artificial intelligence*. AUAI Press, 2004, pp. 487–494.

[7] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning, "Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*. Association for Computational Linguistics, 2009, pp. 248–256.

[8] D. Griffiths and M. Tenenbaum, "Hierarchical topic models and the nested chinese restaurant process," *Advances in neural information processing systems*, vol. 16, p. 17, 2004.

[9] L. Hong and B. D. Davison, "Empirical study of topic modeling in twitter," in *Proceedings of the first workshop on social media analytics*. ACM, 2010, pp. 80–88.

[10] O. Jin, N. N. Liu, K. Zhao, Y. Yu, and Q. Yang, "Transferring topical knowledge from auxiliary long texts for short text clustering," in *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 2011, pp. 775–784.

[11] X. Yan, J. Guo, S. Liu, X. Cheng, and Y. Wang, "Learning topics in short texts by non-negative matrix factorization on term correlation matrix," in *Proceedings of the SIAM International Conference on Data Mining*. SIAM, 2013.

[12] X. Yan, J. Guo, S. Liu, X.-q. Cheng, and Y. Wang, "Clustering short text using ncut-weighted non-negative matrix factorization," in *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, 2012, pp. 2259–2262.

[13] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing twitter and traditional media using topic models," in *Advances in Information Retrieval*. Springer, 2011, pp. 338–349.

[14] T. Lin, W. Tian, Q. Mei, and H. Cheng, "The dual-sparse topic model: mining focused topics and focused terms in short text," in *Proceedings of the 23rd international conference on World wide web*. ACM, 2014, pp. 539–550.

[15] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," in *Proceedings of the 22nd international conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2013, pp. 1445–1456.

[16] Y. Xia, N. Tang, A. Hussain, and E. Cambria, "Discriminative bi-term topic model for headline-based social news clustering," 2015.

[17] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011, pp. 262–272.

[18] Y. Zhao and G. Karypis, "Criterion functions for document clustering: Experiments and analysis," Citeseer, Tech. Rep., 2001.

[19] A. Strehl and J. Ghosh, "Cluster ensembles—a knowledge reuse framework for combining multiple partitions," *The Journal of Machine Learning Research*, vol. 3, pp. 583–617, 2003.

[20] L. Hubert and P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, no. 1, pp. 193–218, 1985.